An Alternative to Traditional Goodness-of-Fit Tests for Discretely Measured Continuous Data

KaDonna C. Randolph and Bill Seaver

Abstract: Traditional goodness-of-fit tests such as the Kolmogorov-Smirnov and χ^2 tests are easily applied to data of the continuous or discrete type, respectively. Occasionally, however, the case arises when continuous data are recorded into discrete categories due to an imprecise measurement system. In this instance, the traditional goodness-of-fit tests may not be wholly applicable because of an unmanageable number of ties in the data, sparse contingency tables, or both; therefore, a flexible alternative to goodness-of-fit tests for discretely measured continuous data is presented. The proposed methodology bootstraps confidence intervals for the difference between selected percentiles of the empirical distribution functions of two samples. Application of the approach is illustrated with a comparison of loblolly pine (*Pinus taeda* L.) tree crown density distributions at the 10th, 25th, 50th, 75th, and 90th percentiles simultaneously. FOR. SCI. 53(5):590–599.

Keywords: bootstrapping, crown density, empirical distribution function, percentiles

A BASIC TASK in scientific research is to determine the extent to which two independent samples differ from one another, and one of the most common ways of doing so is by comparing the means or medians of the two groups. Consider the null hypothesis,

$$H_0: \theta_1 = \theta_2, \quad (1)$$

where θ_i is a measure of location associated with group i ($i = 1, 2$). The null Hypothesis 1 can be tested by calculating a two-sided 100 $(1 - \alpha)$% confidence interval for the difference $\Delta = (\theta_1 - \theta_2)$. If this confidence interval does not include 0, then Hypothesis 1 is rejected at the αth level of significance (Montgomery 1997). Although comparisons based on a single measure of location are often implemented, Wilcox (1995) illustrates how they may miss important differences between two groups, especially if the distributions are heavy tailed or skewed.

To avoid missing important differences that may exist between groups, Wilcox (1995) expanded Hypothesis 1 to include multiple location parameters and tested the expanded hypothesis by calculating simultaneous confidence intervals for $\Delta(x_p) = y_p - x_p$ where p is the pth quantile of groups x and y ($p = 0.1, 0.2, \ldots, 0.9$) and $\Delta(x_p)$ is a measure of how much group x must be shifted so that it is comparable to group y. This function $\Delta(x_p)$ is known as the shift function and was first introduced by Doksum (1974) and Doksum and Sievers (1976).

Another common method of comparing two groups is goodness-of-fit testing. A two-sample goodness-of-fit test is a test of the null hypothesis,

$$H_0: F(x) = H(x) \quad \text{for all } x, \quad (2)$$

where $F(x)$ and (Hx) are the unknown distribution functions associated with the populations being studied. The alternative hypothesis is usually of the general form $F(x) \neq H(x)$ (Reynolds et al. 1988). There are two classical formulations of this hypothesis, the Pearson χ^2 test and the empirical distribution function (EDF) tests.

The Pearson χ^2 test is used with discrete data or continuous data that can be naturally grouped. Let $X_1, X_2, \ldots, X_n$ be a random sample and let $I_1, I_2, \ldots, I_k$ be the partitioned classes for the set of possible values for X. Then the χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i},$$

where f_i is the observed number of observations falling in I_i and $e_i = np_i$ with p_i being the probability of I_i under the null hypothesis (Reynolds et al. 1988). Goodness-of-fit tests for continuous data are based on the EDF. The EDF is defined as follows (Stephens 1986): Let $X_1, X_2, \ldots, X_n$ be a random sample, and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the random sample in rank order. Then,

$$\begin{aligned} F_n(x) &= 0 \quad x < X_{(1)} \\ F_n(x) &= i/n \quad X_{(i)} \leq x < X_{i+1}, \quad i = 1, \ldots, n-1 \\ F_n(x) &= 1 \quad X_{(n)} \leq x. \end{aligned}$$

$F_n(x)$ is a nondecreasing, random function that goes from 0 to 1 in height. It is a step function with steps of height $1/n$ occurring at the sample values (Conover 1999). For any x, $F_n(x)$ is the proportion of observations $\leq x$. $F_n(x)$ is a consistent estimator of $F(x)$, the population cumulative distribution function, and as n goes to infinity, $|F_n(x) - F(x)|$ decreases to 0 with probability 1 (Stephens 1986). The most well-known EDF goodness-of-fit test is based on the Kolmogorov-Smirnov (KS) supremum statistic:

KaDonna C. Randolph, US Forest Service, Southern Research Station FIA, 4700 Old Kingston Pike, Knoxville, TN 37919—Phone: (865) 862-2024; Fax: (865) 862-0262; krandolph@fs.fed.us; and Bill Seaver, Department of Statistics, Operations, and Management Science, College of Business Administration, The University of Tennessee, Knoxville, TN 37996-0562—Phone: (865) 974-6862; wseaver@utk.edu.

Acknowledgments: The authors thank the editor, associate editor, and two anonymous referees for their constructive comments on an earlier draft of this manuscript, which led to substantial improvements in the presentation.

Manuscript received February 8, 2007, accepted April 3, 2007

$$KS = \sup_x |F_n(x) - F(x)| = \max(KS^+, KS^-).$$

Other goodness-of-fit statistics are the Anderson-Darling and Cramér-von Mises statistics,

$$Q = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \psi(x)\, dF(x),$$

where $\psi(x)$ is a function that gives weights to the squared difference $\{F_n(x) - F(x)\}^2$. For the Anderson-Darling statistic $\psi(x)$ equals $[\{F_n(x)\}\{1 - F(x)\}]^{-1}$, and for the Cramér-von Mises statistic $\psi(x)$ equals 1 (Stephens 1986).

In general, the EDF tests are more powerful than the χ^2 tests (Stephens 1974), but they are not well-adapted for truly discrete distributions or for continuous distributions that appear discrete because the data have been grouped (Pettitt and Stephens 1977). When EDF tests are applied to largely discretized data (i.e., data with many ties), the probability of a type I error drops below the nominal level of significance (Wilcox 1997, O'Reilly et al. 2003). χ^2 tests are applicable to data sets with many classes, but use of them with a sparse data set (i.e., several classes with expected cell frequencies <5) may result in poorly approximated test statistics (Agresti 1996). χ^2 approximations for sparse data sets may be improved by combining some of the classes, but doing so often eliminates or obscures important information about the distribution, especially the tails.

Pettitt and Stephens (1977) gave an example of discretized continuous data: angles of inclination of pebbles in moraine deposits, measured to the nearest 5°. Discretized continuous data also occur in image analysis and econometric problems (O'Reilly et al. 2003) and, as will be illustrated, in the analysis of tree crown condition. In reality, all continuous data are subject to some discretization because of imprecise measuring systems but often the impacts of discretization are negligible (Pettitt and Stephens 1977). Regarding the use of discretized continuous data, Schwarz (2006) gives the following rule of thumb: "If the discretization is less than 5% of the typical value, then a discretized continuous variable can be treated as continuous without problems." When discretization is greater than 5% and when the data are too sparse for a χ^2 test, an alternative to the EDF and χ^2 tests is needed. Hence, the purpose of this article is to illustrate a general, large-sample methodology for comparing largely discretized data. The proposed approach improves on the traditional goodness-of-fit methodologies by incorporating more than one measure of location and by accommodating sparse data. These adaptations provide greater flexibility in goodness-of-fit analysis and interpretation. The methodology is similar to the shift function of Wilcox (1995, 1997) but with modifications.

Proposed Methodology

Consider the null hypothesis,

$$H_0: {}^p\Delta = ({}^pX - {}^pY) = 0 \quad \text{for all } p, \tag{3}$$

where ${}^p\Delta$ equals the population parameter describing the differences between the pth percentile of the distribution of populations X and Y, respectively. This hypothesis can be tested using a 100 $(1 - \alpha)$% confidence interval for ${}^p\delta$, which is the sample estimator of ${}^p\Delta$ based on the EDFs of samples taken from X and Y. Hypothesis 3 is rejected if the confidence interval for ${}^p\delta$ does not include 0. Any percentile may be selected and multiple percentiles can be tested simultaneously, but the level of discretization should be the same for the two groups being compared. Wilcox (1995) used nine percentiles ($p = 0.1, 0.2, \ldots, 0.9$) with continuous data. However, with discretized data the level of discretization should be considered to avoid redundant comparisons. If the data are largely discretized so that there are few classes and many observations in each class (i.e., many ties in the data), testing every 5th or 10th percentile may result in redundant comparisons because adjacent percentiles are likely to be equal. Redundant comparisons should be avoided because they reduce the power of the hypothesis test when the probability of a type I error is controlled simultaneously across all comparisons. To illustrate this phenomenon, one might want to primarily test the following percentiles: 5th, 25th, 50th, 75th, and 95th. However, the researcher, on second thought, might also want to test the 20th, 30th, and 35th percentiles as well. These additional percentiles would be redundant and reduce the power of the overall test. Thus, this proposed method is not intended as a broad data snooping option but as a tight hypothesis testing vehicle.

To calculate a confidence interval for ${}^p\delta$, an estimate of standard error is needed. No analytic formula exists for the standard error of differences between percentiles (Mooney and Duval 1993). However, bootstrapping can be used to overcome this limitation. The bootstrap was first introduced by Efron (1979) and provides for the estimation of standard errors for measures of scale and location, such as the mean, variance, skewness, and kurtosis, as well as the median and other percentiles. Hypothesis 3 is tested via bootstrapping with the following generalized algorithm (Carpenter and Bithell 2000):

Designate the first group under consideration as $X = \{x_1, x_2, \ldots, x_n\}$, a sample of n independent observations, then

1. Sample n observations randomly and with replacement from X to obtain a bootstrap data set, denoted X^*.
2. Based on X^*, calculate the p percentile(s) of interest.
3. Repeat steps 1 and 2 i times, generally $i = 1{,}000$ or more, to obtain i estimates of the p percentile(s) of interest.

Next, replace X with Y and repeat steps 1 through 3 the same number of i times. Then calculate ${}^p\delta_i^* = \{{}^pX_i^* - {}^pY_i^*\}$, where ${}^pX_i^*$ and ${}^pY_i^*$ equal the estimate of the pth percentile from the ith bootstrap for groups X and Y, respectively. This calculation establishes the bootstrap distribution of ${}^p\delta$ (designated ${}^p\delta^*$). This process estimates all of the selected percentiles each time a bootstrap resample is made (step 1). For example, if $p = (0.25, 0.50)$ then both percentiles are calculated from the ith bootstrap resample in step 2. This approach is in contrast to that taken by Wilcox (1995), in which only one percentile was estimated from each bootstrap sample (requiring the repetition of steps 1–3 for every pth percentile). Wilcox (1997) estimated all percentiles

from a single bootstrap sample in the comparison of two *dependent* groups and acknowledged that this approach might also be used for independent groups. Although the accuracy of confidence intervals based on estimating all percentiles from a single bootstrap has not been fully investigated, it is our opinion that the difference in accuracy between the two approaches will be inconsequential as long as i is sufficiently large ($\geq$5,000).

An estimate of the standard error of ${}^{p}\delta$ is no longer needed once the bootstrap sampling distribution ${}^{p}\delta^{*}$ is established because the confidence interval for ${}^{p}\delta$ can be determined from ${}^{p}\delta^{*}$ directly. Several options exist for calculating a confidence interval from a bootstrap distribution (Chernick 1999, Carpenter and Bithell 2000), the simplest of which is the percentile (PCTL) method. To illustrate the PCTL method, assume that ${}^{p}\delta^{*}$ consists of 1,000 estimates of ${}^{p}\delta$ and let ${}^{p}\delta^{*}_{(1)}, {}^{p}\delta^{*}_{(2)}, \ldots, {}^{p}\delta^{*}_{(1,000)}$ represent the ordered set so that ${}^{p}\delta^{*}_{(i)} < {}^{p}\delta^{*}_{(j)}$, for $1 \leq i < j \leq 1{,}000$. The lower limit of a two-sided 90% PCTL confidence interval is the 5th percentile of ${}^{p}\delta^{*}$, i.e., ${}^{p}\delta^{*}_{(50)}$, and the upper limit is equal to the 95th percentile, or ${}^{p}\delta^{*}_{(950)}$.

A disadvantage of the PCTL method is that it assumes that ${}^{p}\delta^{*}_{i}$ and ${}^{p}\delta$ are unbiased estimators of ${}^{p}\delta$ and ${}^{p}\Delta$, respectively. To overcome this restriction, the bias corrected (BC) bootstrap confidence interval method makes a correction for median bias by adjusting the upper and lower confidence limit endpoints according to a standardizing transformation that centers the bootstrapped sampling distribution on the point estimator, ${}^{p}\delta$. Typically, the transformation assumes that the bias is normally distributed, although in general, any distributional form may be specified (Mooney and Duval 1993). If the distribution of ${}^{p}\delta^{*}$ is symmetric about ${}^{p}\delta$, i.e., unbiased, then the BC confidence limits are the same as those of the PCTL method. See Chernick (1999) or Mooney and Duval (1993) for further computational details and discussion of other bootstrap confidence interval methods (e.g., the normal approximation and percentile-t methods).

Illustration

Application of this methodology is illustrated with US Forest Service crown condition data. The US Forest Service is charged with reporting the status and trends in forest ecosystem health in the United States. To this end, the US Forest Service Forest Inventory and Analysis (FIA) Program assesses a suite of ecological indicators on a portion of its national network of forest inventory plots (Riitters and Tkacz 2004). One of the ecological indicators of forest health that FIA measures is tree crown density, defined as the amount of crown branches, foliage, and reproductive structures that blocks light visibility through the projected crown outline (US Forest Service 2004). Crown density is visually assessed by a two-person field crew and recorded in 5% increments from 0 to 100.

To be an effective indicator of forest health, tree crown density must be separable into categories of good and poor condition. This requires setting threshold limits to identify the crown densities that signal a decline in tree health. Ideally, this separation should be based on the biological relationship between crown density and another measure of tree vigor (such as diameter increment). Thresholds of this nature are difficult to pinpoint, however, so thresholds based on the statistical distributions are being used until further research is accomplished (Zarnoch et al. 2004).

Statistical thresholds isolate observations in the tails of distributions for designation as either poor or good condition. For example, crown density was initially classified into three categories: poor, 0–20% crown density; moderate, 21–50% crown density; and good, 51–100% crown density (Bechtold 1992). The disadvantage of using such statistically based thresholds is that some observations will be designated as poor even in the absence of a problem (Zarnoch et al. 2004). Furthermore, because of physiological differences, some tree species may be able to tolerate lower levels of crown density than others; hence, one set of thresholds is likely to be insufficient for all species (Randolph 2006). One way to determine whether one set of thresholds is adequate for two groups is to compare the cumulative distribution functions of the groups. If the distributions are not significantly different from one another then a single set of thresholds may be applied to both groups. If the distributions are significantly different then a single set of thresholds is probably inadequate.

To illustrate the proposed methodology, pairwise comparisons were made between the loblolly pine (*Pinus taeda* L.) crown densities in Alabama, North Carolina, and South Carolina. Five levels of p were included in the testing of Hypothesis 3: $p = 0.10, 0.25, 0.50, 0.75$, and 0.90. These five percentiles provide reasonable coverage of the entire distribution and minimize potentially redundant comparisons. Additional percentiles from the tails could have been included (e.g., 5th and 95th percentiles); however, five percentiles were deemed adequate for illustrative purposes.

The experiment-wise type I error rate α was set at 0.10 and 0.20 and a Bonferroni-type correction was implemented so that per comparison confidence intervals were calculated with a confidence level of α/m, where m equals the total number of hypotheses tested (Shaffer 1995). For each state, $i = 5{,}000$ bootstrap resamples were made. The PCTL bootstrap confidence interval method was used to determine the confidence limits for ${}^{p}\delta$. The BC bootstrap confidence interval method was not used because no method to correct the bias of all percentiles simultaneously is available. Null Hypothesis 3 was rejected if any of the confidence intervals in the set excluded 0. The bootstrapping algorithm was performed with SAS software macros available from the SAS Technical Support Web site (SAS Institute, 2004).

Application of the proposed methodology for the comparison of the crown density distributions is warranted because, first of all, a comparison of a single measure of central tendency would be inadequate for describing differences that may occur in the distribution tails that delineate the poorest and best crown conditions. Second, a χ^2 test on the data as presented in Table 1 results in a χ^2 value that may not be valid because a large portion of the cells have expected counts less than the recommended minimum value (typically, 5). Third, if we assume that 40% is a typical crown density value, the level of discretization exceeds the

Table 1. Observed frequencies for loblolly pine crown density measured between 1995 and 1999 in Alabama, North Carolina, and South Carolina

Crown density	Alabama	North Carolina	South Carolina
(%)		.....(no. of trees).....	
0	0	1	0
5	0	1	0
10	0	0	0
15	10	3	0
20	23	1	1
25	62	20	3
30	184	51	3
35	271	84	18
40	247	197	97
45	152	138	156
50	106	95	130
55	47	59	175
60	9	23	112
65	3	15	50
70	2	1	6
75	1	0	7
80	0	0	2
85	0	0	2
90	0	0	0
95	0	0	0
100	0	0	0
Total	1,117	689	762

rule of thumb given by Schwarz (2006), which suggests that the EDF test statistics may be questionable also.

Results

Traditional Goodness-of-Fit Tests

As a point of comparison with the proposed methodology, Hypothesis 2 was tested with the Pearson χ^2 and KS tests for the three pairwise comparisons. Significant χ^2 values ($P < 0.0001$) (Table 2) resulted for all comparisons, but combining of the extreme tail observations into the 25 and 65% categories was required to meet the standard rule of thumb of having expected cell frequency ≥5 for all cells. This aggregation of data affects the tails, which is a key concern for the US Forest Service when identifying healthy and unhealthy trees. Likewise, Hypothesis 2 was rejected ($P < 0.0001$) for the three comparisons when tested with the KS statistic (Table 2). The points of maximum deviation were determined via the KS test to be at crown density values of 35% for the Alabama–North Carolina comparison and 40% for the North Carolina–South Carolina and Alabama–South Carolina comparisons.

Proposed Goodness-of-Fit Test

Hypothesis 3 was rejected at both the 80% and 90% simultaneous confidence levels based on all selected percentiles for the Alabama–South Carolina and North Carolina–South Carolina comparisons. That is, ${}^{p}\delta \neq 0$, $p = 0.10$, 0.25, 0.50, 0.75, and 0.90 (Tables 3 and 4). Examination of the specific confidence intervals shows only one difference between Alabama and North Carolina (${}^{0.9}\delta \neq 0$) (Table 5). The 90th percentile was 5% higher in North Carolina than in Alabama. All percentiles in South Carolina were 10–15% higher than those in Alabama and 5–10% higher than those in North Carolina. There are numerous possibilities for the differences between the loblolly pine crown density distributions. These range from field crew bias in assessment technique to true differences in crown form or crown condition. The specific reasons for the differences are not explored further here; however, they may be investigated in other research efforts.

The conclusions drawn from the results of the proposed methodology are the same as those drawn from the KS and χ^2 tests for the Alabama–South Carolina and North Carolina–South Carolina comparisons. Results from the Alabama–North Carolina comparison are somewhat different, however, and illustrate the impact of tied values on the KS test. For the Alabama–North Carolina comparison, only ${}^{0.9}\delta$ was declared significantly different by way of the proposed methodology. This suggests no differences between the distributions except in the upper tail. On the other hand, the point of maximum deviation determined by the KS test was near the 25th percentile at a crown density value of 35%. This finding suggests differences between the distributions in the lower tail. The difference in conclusions is due to the fact that the KS test is not designed to handle an excessive number of ties in a data set. Notice in Figure 1 that the North Carolina and Alabama distributions are fully separated at the initial occurrence of 35% crown density, but that repeated observations at 35% crown density in North Carolina cause the EDFs to overlap near the 50th percentile. This stacking of the EDF is not considered in the calculation of the KS supremum statistic and results in misleading conclusions about the deviation between the distributions. The proposed methodology accommodates the tied values through bootstrapping and expresses the magnitude of ${}^{p}\delta$ by the confidence interval width (Table 5).

One peculiarity associated with bootstrapping discretized data is evident in the confidence interval for the difference between the 90th percentile of the Alabama–North Carolina crown density distributions where the upper and the lower

Table 2. Results of the Pearson χ^2 and KS goodness-of-fit tests comparing loblolly pine tree crown density in Alabama, North Carolina, and South Carolina

	Original data			Collapsed data[a]			Kolmogorov-Smirnov	
Test	χ^2 value	df	P^b	χ^2 value	df	P^b	D	P^b
AL-NC	152.06	14	<0.0001	138.40	8	<0.0001	0.26	<0.0001
AL-SC	720.61	14	<0.0001	719.28	8	<0.0001	0.55	<0.0001
NC-SC	290.16	16	<0.0001	286.63	8	<0.0001	0.36	<0.0001

[a] Original data were collapsed into fewer categories to achieve expected cell frequencies of ≥5 for all cells.
[b] Probability of observing a larger test statistic under the null hypothesis.

Table 3. Observed percentile estimates for Alabama and South Carolina loblolly pine tree crown density and bootstrap confidence intervals for the difference between observed percentiles (*p*)

	Observed estimate			90% simultaneous[a]		80% simultaneous[b]	
p	AL	SC	Difference (AL-SC)	LCL	UCL	LCL	UCL
			(%)				
0.10	30	40	−10	−15	−10	−15	−10
0.25	35	45	−10	−15	−10	−15	−10
0.50	40	50	−10	−15	−10	−15	−10
0.75	45	55	−10	−15	−10	−15	−10
0.90	50	60	−10	−15	−10	−15	−10

LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $\alpha = 0.04$ per comparison.

Table 4. Observed percentile estimates for North Carolina and South Carolina loblolly pine tree crown density and bootstrap confidence intervals for the difference between observed percentiles (*p*)

	Observed estimate			90% simultaneous[a]		80% simultaneous[b]	
p	NC	SC	Difference (NC-SC)	LCL	UCL	LCL	UCL
			(%)				
0.10	30	40	−10	−10	−5	−10	−5
0.25	40	45	−5	−10	−5	−10	−5
0.50	40	50	−10	−15	−5	−12.5	−5
0.75	50	55	−5	−10	−5	−10	−5
0.90	55	60	−5	−10	−5	−10	−5

LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $\alpha = 0.04$ per comparison.

confidence limits were the same (Table 5). The interval [−5, −5] is a function of both the original data and the bootstrap resamples. The cumulative EDFs for both states are vertical at and around the 90th percentile (Figure 1), which indicates that there is little variation (i.e., many ties) in the crown density values near this percentile. Contrast this finding with the 50th percentile where the cumulative EDF for Alabama is horizontal, which indicates that the 50th percentile occurs near the transition between two crown density classes. When the bootstrap resamples are drawn, the observations for the 90th percentile have very little variation, whereas the observations for the 50th percentile are more heterogeneous. Thus, when the difference at the 90th percentile is calculated, it is nearly constant.

Because a confidence interval with equal endpoints would be extremely unlikely for continuous data, we explored the reasonableness of such an interval for discretized continuous data by perturbing the Alabama and North Carolina data and recalculating the confidence intervals. Recent reports outlining the results from the US Forest Service Forest Health Monitoring Quality Assurance (QA) program discuss the observational differences between two field crews for the crown density indicator and show that there is some variation between the assessments of the same trees by different crews (Pollard and Smith 1999, 2001). Thus, the distribution of observational variation for loblolly pine trees from 6 years of QA assessments (Table 6) was used to perturb the original Alabama and North Carolina crown density distributions. Random numbers were drawn from the uniform (0,1) distribution and the magnitude of variation added to each observation was determined by comparing the random number to the QA cumulative proportion in Table 6. For example, −5 was added to the observation if the random number was >0.09 but ≤0.23.

The resulting proportions of variation added to the crown densities matched well the proportions observed in the

Table 5. Observed percentile estimates for Alabama and North Carolina loblolly pine tree crown density and bootstrap confidence intervals for the difference between observed percentiles (*p*)

	Observed estimate			90% simultaneous[a]		80% simultaneous[b]	
p	AL	NC	Difference (AL-NC)	LCL	UCL	LCL	UCL
			(%)				
0.10	30	30	0	−5	0	−5	0
0.25	35	40	−5	−10	0	−10	0
0.50	40	40	0	−10	0	−10	0
0.75	45	50	−5	−5	0	−5	0
0.90	50	55	−5	−5	−5	−5	−5

LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $\alpha = 0.04$ per comparison.

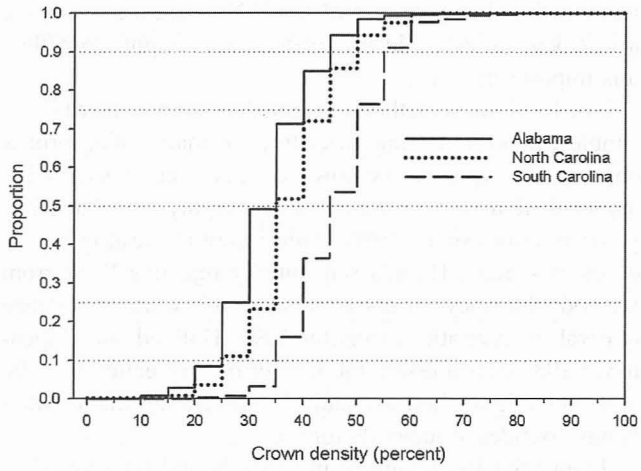

Figure 1. Cumulative empirical distribution functions for Alabama, North Carolina, and South Carolina loblolly pine crown density.

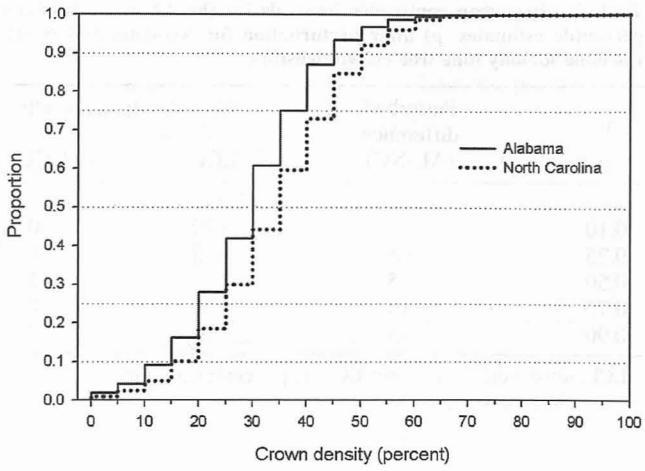

Figure 2. Cumulative empirical distribution functions for Alabama and North Carolina loblolly pine crown density after perturbation of the original data.

original data (Table 6). Truncation to 0 was necessary for seven observations, six in Alabama and one in North Carolina. The addition of the variation smoothed the EDFs (Figure 2); i.e., the steps from one observed crown density to the next were not as high as the original data. Hypothesis 3 was tested with the perturbed data and again was rejected at the 90% confidence level. The perturbation alleviated the zero-width interval for $^{0.9}\delta$ but resulted in an interval of $[-5, -5]$ for $^{0.5}\delta$ (Table 7). Note in Figure 2 that the cumulative EDFs for both states are vertical near the 50th percentile. Recall that this same pattern was evident near the 90th percentile before the addition of observer variation (Figure 1). Thus, even with increased variation, confidence intervals with equal endpoints may result because of the abundance of ties in the data.

Although zero-width confidence intervals may seem to be intuitively incorrect, it is beneficial to remember that the value of each discretized continuous observation is an interval in and of itself. For example, with the exception of 0% crown density, all of the crown density observations represent an interval of 5%: a crown density of 5% includes densities of 1–4%, a crown density of 10% includes densities of 6–10%, and so on (US Forest Service 2004). Thus, there is some inherent width for an interval of $[-5, -5]$ due to the discretized nature of the data.

Discussion

A noteworthy distinction between the proposed alternative methodology and the KS test is that the alternative methodology compares crown density values at selected EDF values (percentiles), whereas the KS test compares all EDF values at each level of crown density. That is, in reference to the EDF plots (Figure 1), the KS test considers vertical differences between the EDFs at each observation along the horizontal, whereas the proposed alternative considers horizontal differences between the EDFs at selected points along the vertical. This characteristic gives the proposed methodology greater flexibility over the traditional goodness-of-fit tests by not only allowing the comparison of percentiles anywhere along the distribution but also by

Table 6. Observed observer variation from FIA QA assessments and proportions of variation added to the observed Alabama and North Carolina crown densities

QA observer variation[a]			Simulated observer variation			
Density variation (%)[b]	Proportion	Cumulative proportion	Proportion (AL)	Cumulative proportion (AL)	Proportion (NC)	Cumulative proportion (NC)
−20	0.01	0.01	0.01	0.01	0.01	0.01
−15	0.03	0.04	0.03	0.04	0.04	0.05
−10	0.05	0.09	0.05	0.09	0.05	0.10
−5	0.14	0.23	0.15	0.24	0.14	0.24
0	0.23	0.46	0.24	0.48	0.22	0.46
5	0.22	0.68	0.21	0.69	0.24	0.70
10	0.16	0.84	0.15	0.84	0.15	0.85
15	0.09	0.93	0.09	0.93	0.08	0.93
20	0.02	0.95	0.02	0.95	0.02	0.95
25	0.03	0.98	0.03	0.98	0.02	0.97
30	0.02	1.00	0.02	1.00	0.03	1.00

[a] From 6 years of QA assessments of loblolly pine in the Southern United States.
[b] Difference between crown density assessments of individual trees by two separate field crews.

Table 7. Bootstrap confidence intervals for the difference between percentile estimates (p) after perturbation for Alabama and North Carolina loblolly pine tree crown densities

p	Perturbed difference (AL-NC)	90% simultaneous CI[a]	
		LCL	UCL
		(%)	
0.10	0	−10	0
0.25	−5	−5	0
0.50	−5	−5	−5
0.75	−10	−10	0
0.90	−5	−10	−5

LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.

providing specific hypothesis testing results for each comparison.

Under the proposed methodology, the individual confidence intervals for ${}^{p}\delta$ reveal the direction and magnitude of the differences for all p included in the hypothesis. Even a confidence interval of 0 width provides useful information about ${}^{p}\delta$. The direction and magnitude of the difference between two distributions can be determined from the KS test results but only for the single point of maximum deviation. Consequently, the KS test does not identify additional differences that might provide important information about the inequality of two distributions.

The capability of the proposed methodology to compare multiple percentiles is an advantage that allows the investigator to focus on areas of the distribution for which differences may have greater consequences. Such might be the case if the data describe conditions of health, as with the crown density data illustrated here. In general, if multiple percentiles are selected across the entire distribution, then there are seven broad ways in which the distributions may differ: at all percentiles; at the lower and middle percentiles, lower and upper percentiles, or middle and upper percentiles; or at the lower, middle, or upper percentiles individually. Each of these combinations will have its own interpretation and importance, depending on the nature of the data and research question under consideration.

Another aspect of the flexibility of the proposed methodology is the ability to control the probability of a type I error simultaneously across all comparisons. In this example, we used a simple (single-step) Bonferroni correction for the α level of significance to construct simultaneous confidence intervals for the set of five percentiles. With the simple Bonferroni correction, each confidence interval was calculated at the $\alpha_i = \alpha/m$ level of significance, where α equals the overall level of type I error and m equals the total number of hypotheses under consideration. This type of correction may be written in the general weighted form as $\alpha_i = w_i\alpha$, where $w_i \geq 0$; $\sum_{i=1}^{m} w_i = 1$. For the simple Bonferroni correction all w_i are equal, i.e., $w_i = 1/m$, although in actuality the α_i may be differentially weighted to allot more of the overall α to the hypotheses considered "most important" (Rosenthal and Rubin 1983, Westfall and Krishen 2001). In our illustration, for example, differences in the lower tail of the EDF (representing the poorest crown conditions) might be considered more important than differences in other segments of the EDFs, and the α_i for the lower percentiles could be given larger weights to reflect this importance.

The Bonferroni method (weighted or nonweighted) is the simplest multiple testing procedure for maintaining proper coverage for type I errors; however, it is a very conservative approach if used with many and/or highly correlated hypothesis tests (Simes 1986). Other multiple testing procedures exist (e.g., Holm's sequentially rejective Bonferroni method), but they do not have straightforward confidence interval interpretations (Shaffer 1995, Holland and Copenhaver 1988). As a result, the Bonferroni correction was the most suitable method available for generating the simultaneous confidence intervals for ${}^{p}\delta$.

Even with its flexibility in analysis and interpretation, there are some limitations to the proposed methodology. One drawback is that when sample sizes are small, the percentile-type confidence intervals, including the BC method, may not maintain true coverage accuracy. Polansky (1999) reports that for a two-sided bootstrap percentile confidence interval for the 50th percentile to maintain true coverage there must be at least 25 observations in the original sample. The number of observations necessary to maintain true coverage for the 75th percentile is 50; 100 observations are needed for the 90th percentile and more than 100 observations for the 95th and 99th percentiles.

The sample sizes used in the illustration were larger than might usually be encountered; therefore, the data sets were reduced in size to examine the sensitivity of the methodology when applied to smaller samples. The original sample sizes were reduced proportionately to 50, 15, and 7.5% of the original sizes. At 7.5% of the original sample size, the sample sizes met Polansky's (1999) recommended number of observations necessary to maintain true coverage at the 75th percentile but not at the 90th percentile. With the smaller sample sizes, the confidence intervals for several ${}^{p}\delta$ widened and affected which ${}^{p}\delta$ were significantly different from 0. For the Alabama–North Carolina comparison, this affected the ultimate conclusion of Hypothesis 3. Given the full data set, Hypothesis 3 for the Alabama–North Carolina comparison was rejected on the basis of the ${}^{0.9}\delta$ confidence interval, which did not include 0 (Table 5). When the sample size was reduced to 15 and 7.5% of the original, the upper confidence limit for ${}^{0.9}\delta$ shifted so that the interval included 0 (Table 8); thus, Hypothesis 3 failed to be rejected. For the North Carolina–South Carolina comparison, the upper confidence limits for ${}^{0.9}\delta$ and ${}^{0.75}\delta$ shifted so that the intervals included 0 (Table 9); however, the overall conclusion of Hypothesis 3 was unchanged (i.e., rejected) because the confidence intervals for the lower percentiles remained significantly different from 0. For the Alabama–South Carolina comparison, shifts in the confidence limits occurred as the sample size was reduced, but there was no change in the individual significance for each ${}^{p}\delta$ nor in the overall conclusion of Hypothesis 3 (Table 10).

The wider confidence intervals observed with the smaller sample sizes were not surprising. As sample sizes decline, there is always less power for hypothesis testing. This was

Table 8. Ninety percent simultaneous[a] bootstrap confidence intervals for the difference between observed percentiles (p) for Alabama and North Carolina loblolly pine tree crown density distributions with reduced sample sizes

	50% sample size[b]		15% sample size[c]		7.5% sample size[d]	
p	LCL	UCL	LCL	UCL	LCL	UCL
	(%)					
0.10	−10	0	−10	0	−10	0
0.25	−10	0	−10	0	−10	0
0.50	−10	0	−10	0	−10	0
0.75	−10	0	−10	0	−10	0
0.90	−5	−5	−10	0	−10	5

Results for the original sample sizes are given in Table 5. LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $n = 559$ for Alabama and $n = 345$ for North Carolina.
[c] $n = 168$ for Alabama and $n = 103$ for North Carolina.
[d] $n = 84$ for Alabama and $n = 52$ for North Carolina.

Table 9. Ninety percent simultaneous[a] bootstrap confidence intervals for the difference between observed percentiles (p) for North Carolina and South Carolina loblolly pine tree crown density distributions with reduced sample sizes

	50% sample size[b]		15% sample size[c]		7.5% sample size[d]	
p	LCL	UCL	LCL	UCL	LCL	UCL
	(%)					
0.10	−10	−5	−15	−5	−15	−2.5
0.25	−10	−5	−10	−5	−15	−2.5
0.50	−15	−5	−15	−5	−15	−5
0.75	−15	−5	−15	−5	−15	0
0.90	−10	−5	−15	−5	−15	0

Results for the original sample sizes are given in Table 4. LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $n = 345$ for North Carolina and $n = 381$ for South Carolina.
[c] $n = 103$ for North Carolina and $n = 114$ for South Carolina.
[d] $n = 52$ for North Carolina and $n = 57$ for South Carolina.

Table 10. Ninety percent simultaneous[a] bootstrap confidence intervals for the difference between observed percentiles (p) for Alabama and South Carolina loblolly pine tree crown density distributions with reduced sample sizes

	50% sample size[b]		15% sample size[c]		7.5% sample size[d]	
p	LCL	UCL	LCL	UCL	LCL	UCL
	(%)					
0.10	−15	−10	−15	−10	−20	−10
0.25	−15	−10	−15	−10	−20	−7.5
0.50	−20	−10	−20	−10	−20	−10
0.75	−15	−10	−20	−10	−20	−10
0.90	−15	−10	−20	−10	−20	−5

Results for the original sample sizes are given in Table 3. LCL, lower confidence limit; UCL, upper confidence limit.
[a] $\alpha = 0.02$ per comparison.
[b] $n = 559$ for Alabama and $n = 381$ for South Carolina.
[c] $n = 168$ for Alabama and $n = 114$ for South Carolina.
[d] $n = 84$ for Alabama and $n = 57$ for South Carolina.

evident with the Alabama–North Carolina comparison particularly. Interval endpoints for the upper percentiles were relatively more unstable than the interval endpoints for the lower percentiles. Thus, caution should be exercised if one is calculating percentile-type confidence intervals for ${}^{p}\delta$ when $p > 0.75$ and sample size is <50. Sample sizes should be >100 if confidence intervals are desired for $p > 0.90$. It is worth noting that sample size is a limitation for calculating confidence intervals for upper-level percentiles in any situation and is not necessarily specific to this application.

Conover (1972), Pettit and Stephens (1977), and Gleser (1985) present KS tests for discrete and discontinuous distributions for small sample sizes ($n < 30$). Although these methods could be extended to larger sample sizes, they are bound by the same limitations as the KS test, namely, as overall goodness-of-fit tests they do not provide the additional specific insight on key percentiles as does the methodology we describe.

A second caveat concerns the occurrence of confidence intervals without width, i.e., when the upper and lower confidence limits are equivalent. Such occurrences are related to the sample size, the discreteness of the data, and the extent to which the two distributions are equivalent. Although highly unusual for continuous data, such intervals were explored by perturbing the data and were shown to be a reasonable possibility for discretized data. This could be a

common occurrence; however, we do not believe it should be considered a major hindrance to the use of the methodology because confidence intervals with no width still provide useful information regarding the degree of difference between two distributions. Nevertheless, investigators considering the proposed methodology should be aware of this possibility and decide whether such occurrences are acceptable for the research question under investigation.

Conclusion

The methodology presented here is a flexible goodness-of-fit test for discretely measured continuous data. The approach is nonparametric and can be applied to distributions of any form, but it is limited to inquiries about the equality of locations for two distributions rather than questions about specific distributional forms, i.e., "are the data normal?" The approach is not intended to take the place of traditional goodness-of-fit tests but rather to provide an alternative method of analyzing discretely measured continuous data. Results of both the traditional and proposed goodness-of-fit tests indicated significant differences between the crown density distributions of the three states; however, the proposed methodology provided greater insight into the nature of the differences. The flexibility not only covers specific percentiles but also allows simulations with different perturbations in a measurement system or with different sample sizes and their effect. In a way, the robustness of this approach is very appealing. On the other hand, this research revealed the need to formulate a multivariate method of handling the bootstrap bias correction method across several percentile estimates. In this case of discontinuous data, such attempts would have been superfluous with no change to the final answers; however, in the case of continuous data, there is a need for such multivariate corrections when there is dependence. Further inquiries into goodness-of-fit testing for discretized data should explore the impact of discretization level and type I error controls (per comparison or simultaneous) on confidence interval widths and coverage accuracy.

Literature Cited

AGRESTI, A. 1996. *An introduction to categorical data analysis.* John Wiley and Sons, New York, NY. 290 p.

BECHTOLD, W.A., W.H. HOFFARD, AND R.L. ANDERSON. 1992. *Summary report: Forest health monitoring in the South, 1991.* US For. Serv. Gen. Tech. Rep. SE-GTR-81. 40 p.

CARPENTER, J., AND J. BITHELL. 2000. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Stat. Med.* 19:1141–1164.

CHERNICK, M.R. 1999. *Bootstrap methods: A practitioner's guide.* John Wiley and Sons, New York, NY. 264 p.

CONOVER, W.J. 1972. A Kolmogorov goodness-of-fit test for discontinuous distributions. *J. Am. Stat. Assoc.* 67(339): 591–596.

CONOVER, W.J. 1999. *Practical nonparametric statistics*, 3rd ed. John Wiley and Sons, New York, NY. 584 p.

DOKSUM, K.A. 1974. Empirical probability plots and statistical inference for nonlinear models in the two sample case. *Ann. Stat.* 2:267–277.

DOKSUM, K.A., AND G.L. SIEVERS. 1976. Plotting with confidence: Graphical comparison of two populations. *Biometrika* 63:421–434.

EFRON, B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 7:1–26.

GLESER, L.J. 1985. Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *J. Am. Stat. Assoc.* 80(392):954–958.

HOLLAND, B.S., AND M.D. COPENHAVER. 1988. Improved Bonferroni-type multiple testing procedures. *Psychol. Bull.* 104(1):145–149.

MONTGOMERY, D.C. 1997. *Design and analysis of experiments*, 4th ed. John Wiley and Sons, New York, NY. 704 p.

MOONEY, C.Z., AND R.D. DUVAL. 1993. *Bootstrapping: A nonparametric approach to statistical inference.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-095. Sage, Newbury Park, CA. 73 p.

O'REILLY, F.J., R. RUEDA, AND M. GARZA.-JINICH. 2003. How important is the effect of rounding in goodness-of-fit. *Commun. Stat.-Simul. C.* 32(3):953–976.

PETTIT, A.N., AND M.A. STEPHENS. 1977. The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* 19(2):205–210.

POLANSKY, A.M. 1999. Upper bounds on the true coverage of bootstrap percentile type confidence intervals. *Am. Stat.* 53(4):362–369.

POLLARD, J.E., AND W. SMITH. 1999 *Forest Health Monitoring 1998 plot component quality assurance report*, vol I. US Forest Service, National Forest Health Monitoring Program, Research Triangle Park, NC.

POLLARD, J.E., AND W. SMITH. 2001. *Forest Health Monitoring 1999 plot component quality assurance report.* US Forest Service, National Forest Health Monitoring Program, Research Triangle Park, NC.

RANDOLPH, K.C. 2006. *Descriptive statistics of tree crown condition in the Southern United States and impacts on data analysis and interpretation.* US For. Serv. Gen. Tech. Rep. SRS-GTR-94. 17 p.

REYNOLDS, M.R., JR., T.E. BURK, AND W. HUANG. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. *For. Sci.* 34(2):373–399.

RIITTERS, K., AND B. TKACZ. 2004. The US Forest Health Monitoring Program. P. 669–683 in *Environmental monitoring*, Wiersma, B. (ed.). CRC Press, Boca Raton, FL.

ROSENTHAL, R., AND D.B. RUBIN. 1983. Ensemble-adjusted p values. *Psychol. Bull.* 94(3):540–541.

SAS INSTITUTE. 2004. *Jackknife and bootstrap analyses.* Available online at ftp.sas.com/techsup/download/stat/jackboot.html; last accessed Aug. 21, 2007.

SCHWARZ, C.J. 2006. *Types and roles of data.* Available online at www.math.sfu.ca/~cschwarz/Stat-201/Handouts/node6.html; last accessed Aug. 21, 2007.

SHAFFER, J.P. 1995. Multiple hypothesis testing. *Annu. Rev. Psychol.* 46:561–584.

SIMES, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73(3):751–754.

STEPHENS, M.A. 1974. EDF Statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* 69(347):730–737.

STEPHENS, M.A. 1986. Tests based on EDF statistics. P. 97–193 in *Goodness-of-fit techniques*, D'Agostino, R.B., and M.A. Stephens (eds.). Marcel Dekker, New York, NY. 560 p.

US FOREST SERVICE. 2004. *Forest inventory and analysis field methods for phase 3 measurements, version 2.0, section 12.0 crowns: Measurements and sampling.* Available online at www.fia.fs.fed.us/library/field-guides-methods-proc/docs/

p3_2-0_sec12_3_04.pdf; last accessed Dec. 8, 2005.

WESTFALL, P.H., AND A. KRISHEN. 2001. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Stat. Planning Inference* 99:25–40.

WILCOX, R.R. 1995. Comparing two independent groups via multiple quantiles. *Statistician* 44(1):91–99.

WILCOX, R.R. 1997. *The shift function.* P. 93–105 in *Introduction to robust estimation and hypothesis testing.* Academic Press, San Diego, CA. 296 p.

ZARNOCH, S.J., W.A. BECHTOLD., AND K.W. STOLTE. 2004. Crown condition as an indicator of forest health. *Can. J. For. Res.* 34:1057–1070.